KU LEUVEN

FACULTY OF ECONOMICS AND BUSINESS

Non linear mixed models for predictive modelling in actuarial science

Katrien Antonio, Yanwei Zhang

AFI_1384

Non linear mixed models for predictive modelling in actuarial science

Katrien Antonio * Yanwei Zhang †

November 24, 2013

Preview of the Chapter. We start with a discussion of model families for multilevel data outside the Gaussian framework. We continue with Generalized Linear Mixed Models ([GLMMs]), which enable generalized linear modeling with multilevel data. The Chapter includes highlights of estimation techniques for GLMMs, in the frequentist as well as Bayesian context. We continue with a discussion of Non Linear Mixed Models ([NLMMs]). The Chapter concludes with an extensive case study using a selection of `R` packages for GLMMs.

1 Introduction

Chapter XXX (Section XXX) motivates predictive modeling in actuarial science (and in many other statistical disciplines) when data structures go beyond the cross–sectional design. Mixed (or: multilevel) models are statistical models suitable for the analysis of data structured in nested (i.e. *hierarchical*) or non–nested (i.e. cross–classified, *next to* each other instead of hierachically nested) **clusters or levels**. While the focus in Chapter XXX is on *linear* mixed models, we will now extend the idea of mixed modeling to outcomes with a distribution from the exponential family (as in the Chapter on Generalized Linear Models [GLMs]) and to mixed models which release the concept of linear predictors. The first extension leads to the family of Generalized Linear Mixed Models ([GLMMs]) and the latter creates Non–Linear Mixed Models ([NLMMs]). The use of mixed models for predictive modeling with multilevel data is motivated extensively in Chapter XXX . These motivations also apply here. We focus in this Chapter on the formulation, calibration and interpretation of mixed models for non normal outcomes and non linear modeling problems. Estimation, inference and prediction with a range of numerical techniques are discussed. Readers who are not interested in technicalities of estimation techniques can skip Sections 3.3.1 to 3.3.3.

[Reference to Chapter on linear mixed models, examples section.]

[Reference to Chapter on linear mixed models, Introduction.]

*University of Amsterdam and KU Leuven (Belgium), email: k.antonio@uva.nl
†University of Southern California, email: actuary_zhang@hotmail.com

2 Model families for multilevel non–Gaussian data

Section XXX (Chapter XXX) explains the connection between the marginal and hierarchical interpretation of a linear mixed model ([LMM]). This feature is a consequence of the properties of the multivariate normal distribution, but it will no longer exist when outcomes are of non–Gaussian type. Thus, with outcomes of non–Gaussian type we explicitly distinguish so–called *marginal* (cfr. infra) versus *random effects* models for clustered (or: multilevel) non–normal data. Molenberghs and Verbeke (2005) distinguish three families of models for handling non–Gaussian clustered data: ***marginal***, ***conditional*** and ***subject–specific*** models. Generalized Estimating Equations ([GEEs]) (see Liang and Zeger (1986)) are a well–known computational tool for ***marginal models***. With GEEs the marginal mean $\boldsymbol{\mu} = E[\boldsymbol{y}] = g^{-1}(\boldsymbol{X\beta})$ should be correctly specified, in combination with a working assumption for the dependence structure. $g(.)$ is the link function introduced in Chapter XXX . Applications of GEEs in actuarial predictive modeling are in Purcaru et al. (2004) and Denuit et al. (2007), but are not covered here. The class of ***conditional models*** is a second group of models where $\boldsymbol{y}$ is modeled conditional upon (a subset of) the other outcomes. We will not discuss these models here. Our focus – from Section 3 on – is on ***subject or cluster–specific models***, more specifically on generalized and non linear mixed models ([GLMMs] and [NLMMs]) incorporating random, subject or cluster–specific effects.

[Reference to Chapter on GLMs.]

3 Generalized Linear Mixed Models

3.1 Generalized linear models

Generalized Linear Models ([GLMs]) have numerous applications in actuarial science, ranging from ratemaking over loss reserving to mortality modeling. See Haberman and Renshaw (1996) for an overview. Chapter XXX of this book explains in detail the use of GLMs with cross–sectional data. A GLM is a regression model specified for a distribution from the exponential family. A member of this family has a density of the form

[Reference to the chapter on GLMs.]

$$f_Y(y) = \exp\left(\frac{y\theta - \psi(\theta)}{\phi} + c(y,\phi)\right). \tag{1}$$

$\psi(.)$ and $c(.)$ are known functions, θ is the natural and ϕ the scale parameter. Using vector notation the following relations hold

$$\boldsymbol{\mu} = \mathrm{E}[\boldsymbol{y}] = \psi^{'}(\boldsymbol{\theta}) \quad \text{and} \quad \mathrm{Var}[\boldsymbol{y}] = \phi\psi^{''}(\boldsymbol{\theta}) = \phi V(\boldsymbol{\mu}), \tag{2}$$

where derivatives are with respect to $\boldsymbol{\theta}$ and $V(.)$ is the so–called variance function. The latter function captures the relationship between the mean and variance of $\boldsymbol{y}$. GLMs provide a way around transforming data, by specifying a linear predictor for a transformation of the mean

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{X\beta}, \tag{3}$$

with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{'}$ the vector of regression parameters and $\boldsymbol{X}$ $(m \times p)$ the design matrix. g is the link function and $\boldsymbol{\eta}$ the so–called linear predictor. Estimates for $\boldsymbol{\beta}$ follow by solving the maximum likelihood equations with an iterative numerical technique (such

as Newton-Raphson). Likelihood ratio and Wald tests are available for inference purposes. If the scale parameter ϕ is unknown, we estimate it by maximum likelihood or by dividing the deviance or Pearson's chi–square statistic by its degrees of freedom.

3.2 Extending GLMs with random effects

GLMMs extend GLMs by adding random effects $\boldsymbol{Zu}$ to the linear predictor $\boldsymbol{X\beta}$. Motivations for this extension are similar to those in Section XXX from Chapter XXX: the random effects enable cluster–specific prediction, they allow for heterogeneity between clusters and structure correlation within clusters. Conditional on a q–dimensional vector $\boldsymbol{u}_i$ of random effects for cluster i, GLMM assumptions for the jth response on cluster or subject i, y_{ij}, are

$$\begin{aligned} y_{ij}|\boldsymbol{u}_i &\sim f_{Y_{ij}|\boldsymbol{u}_i}(y_{ij}|\boldsymbol{u}_i) \\ f_{Y_{ij}|\boldsymbol{u}_i}(y_{ij}|\boldsymbol{u}_i) &= \exp\left(\frac{y_{ij}\theta_{ij} - \psi(\theta_{ij})}{\phi} - c(y_{ij},\phi)\right) \\ \boldsymbol{u}_i &\sim f_U(\boldsymbol{u}_i), \end{aligned} \tag{4}$$

with $\boldsymbol{u}_i$ independent among clusters i. The following conditional relations hold

$$\mu_{ij} = E[y_{ij}|\boldsymbol{u}_i] = \psi^{'}(\theta_{ij}) \quad \text{and} \quad \text{Var}[y_{ij}|\boldsymbol{u}_i] = \phi\psi^{''}(\theta_{ij}) = \phi V(\mu_{ij}). \tag{5}$$

A transformation of the mean μ_{ij} is linear in both the fixed ($\boldsymbol{\beta}$) and random effects ($\boldsymbol{u}_i$) parameter vectors

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{'}\boldsymbol{u}_i, \tag{6}$$

with $\boldsymbol{u}_i$ the vector of random effects for cluster i, $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ the p and q dimensional vectors of known covariates corresponding with the fixed and random effects, respectively. A distributional assumption for the random effects vector $\boldsymbol{u}_i$, say $f_{\boldsymbol{U}}(\boldsymbol{u}_i)$, completes the specification of a GLMM. Most applications use normally distributed random effects, but other distributional assumptions are possible.

The model assumptions in (4), (5) and (6) imply the following specifications for marginal mean and variance

$$\begin{aligned} E[y_{ij}] &= E[E[y_{ij}|\boldsymbol{u}_i]] = E[g^{-1}(\boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{'}\boldsymbol{u}_i)] \\ \text{Var}(y_{ij}) &= \text{Var}(E[y_{ij}|\boldsymbol{u}_i]) + E[\text{Var}(y_{ij}|\boldsymbol{u}_i)] \\ &= \text{Var}(\mu_{ij}) + E[\phi V(\mu_{ij})] \\ &= \text{Var}(g^{-1}[\boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{'}\boldsymbol{u}_i]) + E[\phi V(g^{-1}[\boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{'}\boldsymbol{u}_i])]. \end{aligned} \tag{7}$$

In general, simplification of these expressions is not possible. The GLMM regression parameters $\boldsymbol{\beta}$ do not have a marginal interpretation; they express the effect of a set of covariates on the response, conditional on the random effects $\boldsymbol{u}_i$. Indeed, $\text{E}[y_{ij}] = \text{E}[\text{E}[y_{ij}|\boldsymbol{u}]] = \text{E}[g^{-1}(\boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{'}\boldsymbol{u}_i)] \neq g^{-1}(\boldsymbol{x}_{ij}^{'}\boldsymbol{\beta})$. Illustration 1 shows explicit calculation of marginal mean, variance and covariance within a Poisson GLMM.

Illustration 1 (A Poisson GLMM)**.** *Conditional on a random intercept* $u_i \sim N(0, \sigma^2)$*,* y_{ij} *is Poisson distributed with* $\mu_{ij} = E[y_{ij}|u_i] = \exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta} + u_i)$*. Thus, the link function* g *is the logarithm. The corresponding likelihood is*

$$L(\boldsymbol{\beta}, \sigma|\boldsymbol{y}) = \prod_{i=1}^{m} \int_{-\infty}^{+\infty} \left(\prod_{j=1}^{n_i} \frac{\mu_{ij} e^{-\mu_{ij}}}{y_{ij}!} \right) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2} u_i^2} du_i. \tag{8}$$

Straightforward calculations using mean and variance of the lognormal distribution show

$$\begin{aligned} E(y_{ij}) &= E(E(y_{ij}|u_i)) = E(\exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta} + u_i)) \\ &= \exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta}) \exp(\sigma^2/2) \qquad (9) \\ \textit{and} & \\ Var(y_{ij}) &= Var(E(y_{ij}|u_i)) + E(Var(y_{ij}|u_i)) \\ &= E(y_{ij})(\exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta})[\exp(3\sigma^2/2) - \exp(\sigma^2/2)] + 1), \qquad (10) \\ \textit{and} & \\ Cov(y_{ij}, y_{ik}) &= Cov(E(y_{ij}|u_i), E(y_{ik}|u_i)) + E(Cov(y_{ij}, y_{ik}|u_i)) \qquad (j \neq k) \\ &= \exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta}) \exp(\boldsymbol{x}_{ik}'\boldsymbol{\beta})(\exp(2\sigma^2) - \exp(\sigma^2)). \qquad (11) \end{aligned}$$

The expression in round parentheses in (10) *is always greater than 1. Thus, although* $y_{ij}|u_i$ *follows a regular Poisson distribution, the marginal distribution of* y_{ij} *is over–dispersed. According to* (11)*, due to the random intercept, observations on the same subject are no longer independent, as is desirable for clustered data. Actuarial literature on ratemaking (see e.g. Denuit et al. (2007) and Antonio and Valdez (2012)) often uses a slightly modified version of the normality assumption, namely* $u_i \sim N(-\frac{\sigma^2}{2}, \sigma^2)$*. This leads to*

$$\begin{aligned} E[y_{ij}] &= E[E[y_{ij}|u_i]] = \exp(\boldsymbol{x}_i'\boldsymbol{\beta} - \frac{\sigma^2}{2} + \frac{\sigma^2}{2}) \\ &= \exp(\boldsymbol{x}_i'\boldsymbol{\beta}), \\ E[y_{ij}|u_i] &= \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + u_i). \qquad (12) \end{aligned}$$

In actuarial parlance, the so–called ***a priori*** *premium (*$E[y_{ij}]$*), specified as* $\exp(\boldsymbol{x}_i'\boldsymbol{\beta})$*, uses only a priori measurable risk factors (like gender, age, car capacity, ...). It is the marginal mean of* y_{ij} *and is therefore correct on average. The* ***a posteriori*** *correction factor,* $\exp(u_i)$*, adjusts the a priori tariff based on the observed claim history of the insured. We estimate this factor by predicting* u_i*.*

Illustration 2 (An illustration of shrinking)**.** *We consider a claim frequency model using the auto claim data from Yip and Yau (2005), where we specify a log-linear Poisson model with* Jobclass *as random effect. In particular, we are interested at how the estimate for each job class level differs between the mixed model and the GLM where* Jobclass *enters as a factor fixed effect. This is the difference between a partial pooling approach (with mixed models) and the 'no pooling' approach (with cluster specific intercepts), see our discussion in Chapter XXX. Figure 1 shows such a comparison on the estimation of job class levels. The horizontal dotted line corresponds to the estimated intercept from the*

mixed model and represents the average effect for all job categories because all the random effects have zero means. That is, it is roughly the estimate when all job categories are pooled together. On the other hand, the estimates from the generalized linear model (the red points) can be viewed as the individual estimate for each job class level ignoring the other levels - indeed, fitting a GLM with only the job class as a predictor is equivalent to fitting 8 separate GLMs on each subset of data with a unique job class because of the orthogonal design matrix corresponding to the job class. We see that the mixed model (the green triangle) shrinks the separate estimates from the GLM toward the pooled group-level estimate across all the job classes. The shrinkage is most significant for Lawyer, Professional and Student. Therefore, the generalized linear mixed model captures the core insight of the credibility models, where the estimates from the mixed models are can be viewed as the weighted average between the pooled group-level estimate and the separate individual estimates. As a result, the mixed model produces less extreme estimates while still accounting for the heterogeneity across the various levels.

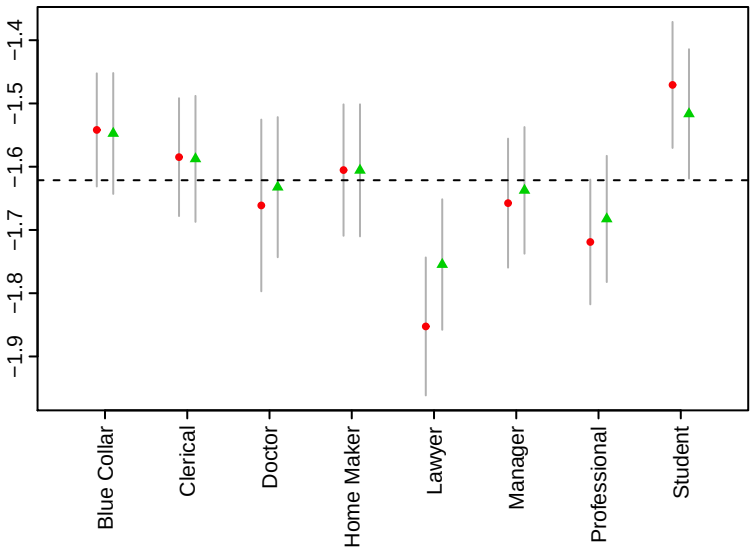

Figure 1: *The job class estimates from the generalized linear model (•) and the Poisson mixed models (△) in the auto insurance frequency model. The horizontal line is the average estimate for all job classes, and the vertical lines show the uncertainty intervals based on ± one standard errors.*

3.3 Estimation

Using the model specifications in (4) it is straightforward to write down the likelihood of the corresponding GLMM

$$L(\boldsymbol{\beta}, \boldsymbol{D}|\boldsymbol{y}) \quad = \quad \int f_{\boldsymbol{Y}|\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{u}) f_{\boldsymbol{U}}(\boldsymbol{u}) d\boldsymbol{u}, \tag{13}$$

where the integral goes over the random effects vector $\boldsymbol{u}$ (with covariance matrix $\boldsymbol{D}$). The presence of the integral in (13) hinders maximum likelihood estimation and prohibits explicit expressions for estimators and predictors, like those derived for LMMs. Only

so–called *conjugate* distributional specifications lead to a closed–form solution in (13); a normal distribution for the response, combined with normally distributed random effects (as with LMMs) being one example. More general model assumptions require approximate techniques to estimate $\boldsymbol{\beta}$, $\boldsymbol{D}$ and predict the random effect for cluster i, $\boldsymbol{u}_i$. As in Molenberghs and Verbeke (2005) we distinguish three approaches to tackle this problem: approximating the integrand, approximating the data and approximating the integral (through numerical integration). Having Pinheiro and Bates (2000), McCulloch and Searle (2001) (Chapters 8 and 10) and Tuerlinckx et al. (2006) as main references, we discuss below some highlights of these methods. This discussion will help readers to understand the differences between e.g. different `R` packages available for data analysis with GLMMs (as demonstrated in Section 6.1). Section 3.4 presents pros and cons of the techniques mentioned in 3.3.1, 3.3.2 and 3.3.3, as well as references to other techniques (not discussed here). We postpone a discussion of a Bayesian approach to Section 5.

3.3.1 Approximating the likelihood: the Laplace method

The Laplace method (see Tierny and Kadane (1986)) approximates integrals of the form

$$\int e^{h(\boldsymbol{u})} d\boldsymbol{u}. \tag{14}$$

for some function h of a q–dimensional vector $\boldsymbol{u}$. The method relies on a second–order Taylor expansion of $h(\boldsymbol{u})$ around its maximum $\hat{\boldsymbol{u}}$

$$h(\boldsymbol{u}) \approx h(\hat{\boldsymbol{u}}) + \frac{1}{2}(\boldsymbol{u}-\hat{\boldsymbol{u}})' h''(\hat{\boldsymbol{u}})(\boldsymbol{u}-\hat{\boldsymbol{u}}), \tag{15}$$

with

$$\frac{\partial h(\boldsymbol{u})}{\partial u}|_{\boldsymbol{u}=\hat{\boldsymbol{u}}} = \boldsymbol{0}, \tag{16}$$

and $h''(\hat{\boldsymbol{u}}) = \frac{\partial^2 h(\boldsymbol{u})}{\partial \boldsymbol{u} \partial \boldsymbol{u}'}\Big|_{\boldsymbol{u}=\hat{\boldsymbol{u}}}$ the matrix with second order derivatives of h, evaluated at $\hat{\boldsymbol{u}}$. We replace $h(\boldsymbol{u})$ with the approximation from (15)

$$\int e^{h(\boldsymbol{u})} d\boldsymbol{u} \approx \int e^{h(\hat{\boldsymbol{u}}) + \frac{1}{2}(\boldsymbol{u}-\hat{\boldsymbol{u}})' h''(\hat{\boldsymbol{u}})(\boldsymbol{u}-\hat{\boldsymbol{u}})} d\boldsymbol{u}. \tag{17}$$

The right hand side of (17) is proportional to the integral over a Gaussian density function with mean $\hat{\boldsymbol{u}}$ and variance $(-h''(\hat{\boldsymbol{u}}))^{-1}$). Thus, it can be easily evaluated as

$$\int e^{h(\boldsymbol{u})} d\boldsymbol{u} \approx (2\pi)^{q/2} \left| -h''(\hat{\boldsymbol{u}}) \right|^{-1/2} e^{h(\hat{\boldsymbol{u}})}. \tag{18}$$

This technique is readily available to approximate the likelihood in a GLMM (see Breslow and Clayton (1993) and McCulloch and Searle (2001), among other references)

$$\begin{aligned} \ell &= \log \int f_{\boldsymbol{Y}|\boldsymbol{U}}(\boldsymbol{y}|\boldsymbol{u}) f_{\boldsymbol{U}}(\boldsymbol{u}) d\boldsymbol{u} \\ &= \log \int e^{\log f_{\boldsymbol{Y}|\boldsymbol{U}}(\boldsymbol{y}|\boldsymbol{u}) + \log f_{\boldsymbol{U}}(\boldsymbol{u})} d\boldsymbol{u} \\ &= \log \int e^{h(\boldsymbol{u})} d\boldsymbol{u}, \end{aligned} \tag{19}$$

with $h(\boldsymbol{u}) := \log f_{\boldsymbol{Y}|\boldsymbol{U}}(\boldsymbol{y}|\boldsymbol{u}) + \log f_{\boldsymbol{U}}(\boldsymbol{u}) = \log f_{\boldsymbol{Y}|\boldsymbol{U}}(\boldsymbol{y}|\boldsymbol{u}) - \frac{1}{2}\boldsymbol{u}^{'}\boldsymbol{D}^{-1}\boldsymbol{u} - \frac{q}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{D}|$. (16) should be solved numerically and requires

$$\begin{aligned} \frac{\partial h(\boldsymbol{u})}{\partial \boldsymbol{u}} = \frac{\partial \log f_{\boldsymbol{Y}|\boldsymbol{U}}(\boldsymbol{y}|\boldsymbol{u})}{\partial \boldsymbol{u}} - \boldsymbol{D}^{-1}\boldsymbol{u} &= \boldsymbol{0} \\ &\Updownarrow \\ \frac{1}{\phi}\boldsymbol{Z}^{'}\boldsymbol{W}\boldsymbol{\Delta}(\boldsymbol{y}-\boldsymbol{\mu}) - \boldsymbol{D}^{-1}\boldsymbol{u} &= \boldsymbol{0}, \end{aligned} \tag{20}$$

where $\boldsymbol{W}$ and $\boldsymbol{\Delta}$ are diagonal matrices with elements $[V(\mu_i)(g^{'}(\mu_i))^2]^{-1}$ and $g^{'}(\mu_i)$, respectively [1]. Hereby $g(\mu_i)$ and $V(\mu_i)$ are the mean and variance of y_i, conditional on $\boldsymbol{u}_i$, as introduced in (2).

The matrix of second order derivatives is (see (18))

$$\frac{\partial^2 h(\boldsymbol{u})}{\partial \boldsymbol{u}\partial \boldsymbol{u}^{'}} = -\frac{1}{\phi}\boldsymbol{Z}^{'}\boldsymbol{W}\boldsymbol{\Delta}\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{u}^{'}} + \frac{1}{\phi}\boldsymbol{Z}^{'}\frac{\partial \boldsymbol{W}\boldsymbol{\Delta}}{\partial \boldsymbol{u}^{'}}(\boldsymbol{y}-\boldsymbol{\mu}) - \boldsymbol{D}^{-1}. \tag{21}$$

The random vector corresponding with the second term in this expression has expectation zero, with respect to $f_{\boldsymbol{Y}|\boldsymbol{U}}(\boldsymbol{y}|\boldsymbol{u})$, and will be ignored. Therefore,

$$\begin{aligned} -\frac{\partial^2 h(\boldsymbol{u})}{\partial \boldsymbol{u}\partial \boldsymbol{u}^{'}} &\approx \frac{1}{\phi}\boldsymbol{Z}^{'}\boldsymbol{W}\boldsymbol{\Delta}\boldsymbol{\Delta}^{-1}\boldsymbol{Z} + \boldsymbol{D}^{-1} \\ &= \left(\frac{1}{\phi}\boldsymbol{Z}^{'}\boldsymbol{W}\boldsymbol{Z}\boldsymbol{D} + \boldsymbol{I}\right)\boldsymbol{D}^{-1}. \end{aligned} \tag{22}$$

Using this expression an approximation to the log–likelihood in (19) follows

$$\begin{aligned} \ell &\approx \log f_{\boldsymbol{Y}|\boldsymbol{U}}(\boldsymbol{y}|\hat{\boldsymbol{u}}) - \frac{1}{2}\hat{\boldsymbol{u}}^{'}\boldsymbol{D}^{-1}\hat{\boldsymbol{u}} - \frac{q}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{D}| \\ &\quad + \frac{q}{2}\log 2\pi - \frac{1}{2}\log|(\boldsymbol{Z}^{'}\boldsymbol{W}\boldsymbol{Z}\boldsymbol{D}/\phi + \boldsymbol{I})\boldsymbol{D}^{-1}| \\ &= \log f_{\boldsymbol{Y}|\boldsymbol{U}}(\boldsymbol{y}|\hat{\boldsymbol{u}}) - \frac{1}{2}\hat{\boldsymbol{u}}^{'}\boldsymbol{D}^{-1}\hat{\boldsymbol{u}} + \frac{1}{2}\log|\boldsymbol{Z}^{'}\boldsymbol{W}\boldsymbol{Z}\boldsymbol{D}/\phi + \boldsymbol{I}|. \end{aligned} \tag{23}$$

This expression should be maximized with respect to $\boldsymbol{\beta}$. McCulloch and Searle (2001) assume $\boldsymbol{W}$ is not changing a lot as a function of $\boldsymbol{\beta}$, the last term can be ignored [2] and

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \frac{1}{\phi}\boldsymbol{X}^{'}\boldsymbol{W}\boldsymbol{\Delta}(\boldsymbol{y}-\boldsymbol{\mu}). \tag{24}$$

[1] Derivations are similar to those in Chapter XXX on GLMs, and basically go as follows:

$$\begin{aligned} \frac{\partial \log f_{\boldsymbol{Y}|\boldsymbol{U}}(\boldsymbol{y}|\boldsymbol{u})}{\partial \boldsymbol{u}} &= \frac{1}{\phi}\sum_i \left(y_i\frac{\partial \theta_i}{\partial \boldsymbol{u}} - \frac{\partial \psi(\theta_i)}{\partial \theta_i}\frac{\partial \theta_i}{\partial \boldsymbol{u}}\right) \\ &= \frac{1}{\phi}\sum_i (y_i-\mu_i)\frac{1}{V(\mu_i)}\frac{1}{g^{'}(\mu_i)}\boldsymbol{z}_i^{'}. \end{aligned}$$

[2] However, the `lme4` package in `R` does not ignore the last term in this expression, see `http://lme4.r-forge.r-project.org/book/`.

Therefore, the following set of equations has to be solved simultaneously with respect to $\boldsymbol{\beta}$ and $\boldsymbol{u}$ (using a numerical optimization method)

$$\begin{aligned} \frac{1}{\phi}\boldsymbol{X}^{'}\boldsymbol{W}\boldsymbol{\Delta}(\boldsymbol{y}-\boldsymbol{\mu}) &= \boldsymbol{0} \\ \frac{1}{\phi}\boldsymbol{Z}^{'}\boldsymbol{W}\boldsymbol{\Delta}(\boldsymbol{y}-\boldsymbol{\mu}) &= \boldsymbol{D}^{-1}\boldsymbol{u}. \end{aligned} \tag{25}$$

This set of equations also arises by jointly maximizing (with respect to $\boldsymbol{\beta}$ and $\boldsymbol{u}$)

$$\log f_{\boldsymbol{Y}|\boldsymbol{U}}(\boldsymbol{y}|\boldsymbol{u}) - \frac{1}{2}\boldsymbol{u}^{'}\boldsymbol{D}^{-1}\boldsymbol{u}, \tag{26}$$

which is a quasi–likelihood term, $f_{\boldsymbol{Y}|\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{u})$, augmented with a penalty term, $\boldsymbol{u}^{'}\boldsymbol{D}\boldsymbol{u}$. Hence, the name Penalized Quasi–Likelihood (PQL) for (26). Breslow and Clayton (1993) present a Fisher scoring algorithm, and its connection with Henderson's mixed model equations for simultaneous solution of the set of equations in (25). This approach is discussed in the next section.

3.3.2 Approximating the data: pseudo–likelihood (PL)

Wolfinger and O'Connell (1993) develop pseudo–likelihood ([PL]) (or restricted pseudo–likelihood, [REPL]) in the context of GLMMs. This approach generalizes the idea of a ***working variate***, introduced for MLE with GLMs (see Chapter XXX), to the case of GLMMs (also see Breslow and Clayton (1993) and McCulloch and Searle (2001)). In the context of GLMs Nelder and Wedderburn (1972) define a working variate t_i as follows

$$\begin{aligned} t_i &= g(\mu_i) + g^{'}(\mu_i)(y_i-\mu_i) \\ &= \boldsymbol{x}_i^{'}\boldsymbol{\beta} + g^{'}(\mu_i)(y_i-\mu_i). \end{aligned} \tag{27}$$

Estimates of $\boldsymbol{\beta}$ follow from iteratively fitting a weighted linear regression of $\boldsymbol{t}$ on $\boldsymbol{X}$, until convergence of the estimates. In a GLMM we generalize the notion of a working variate t_i as follows

$$t_i = \boldsymbol{x}_i^{'}\boldsymbol{\beta} + \boldsymbol{z}_i^{'}\boldsymbol{u} + g^{'}(\mu_i)(y_i-\mu_i). \tag{28}$$

This is a first order Taylor expansion of $g(y_i)$ around the conditional mean μ_i. In matrix notation the vector of working variates, $\boldsymbol{t}$, becomes

$$\boldsymbol{t} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\Delta}(\boldsymbol{y}-\boldsymbol{\mu}), \tag{29}$$

with $\boldsymbol{\Delta}$ a diagonal matrix with entries $g^{'}(\mu_i)$. Calculating the variance of $\boldsymbol{t}$ is complicated because of the dependence of $\boldsymbol{\Delta}$ on $\boldsymbol{\mu}$ (and therefore on the random vector $\boldsymbol{u}$). A simplification is possible by replacing $\boldsymbol{\mu}$ with $\hat{\boldsymbol{\mu}}$ in the variance matrix (see Wolfinger and O'Connell (1993)). Consequently,

$$\begin{aligned} \mathrm{Var}(\boldsymbol{t}) &= \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^{'} + \boldsymbol{\Delta}_{\hat{\mu}}\mathrm{Var}(\boldsymbol{Y}-\boldsymbol{\mu})_{\hat{\mu}}\boldsymbol{\Delta}_{\hat{\mu}} \\ &:= \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^{'} + \boldsymbol{\Sigma}_{\hat{\boldsymbol{\mu}}}. \end{aligned} \tag{30}$$

The working variate $\boldsymbol{t}$ approximately follows a linear mixed model (as in Chapter XXX), with design matrices $\boldsymbol{X}$ (fixed effects), $\boldsymbol{Z}$ (random effects), $\boldsymbol{D}$ the covariance matrix of the random effects and $\boldsymbol{\Sigma}$ the covariance matrix of the error terms. In this LMM it is straightforward to estimate $\boldsymbol{\beta}$, $\boldsymbol{u}$ and the unknown variance components. Therefore, the pseudo–likelihood algorithm goes as follows. Starting from initial estimates of $\boldsymbol{\beta}$, $\boldsymbol{u}$ and the variance components, the working variates in (29) are evaluated. Consequently, using LMM methodology, updated estimates follow from (29) and (30). These steps are repeated until convergence of the estimates.

3.3.3 Approximating the integral: numerical integration techniques

Approximating the integral in (13) with a so–called (adaptive) quadrature rule for numerical integration is based upon Liu and Pierce (1994). For ease of explanation we consider below the case of a one–dimensional integral. The case with multidimensional integrals is documented in Tuerlinckx et al. (2006).

Non-adaptive Gauss–Hermite quadrature. *Non-adaptive* Gauss-Hermite quadrature approximates an integral of the form

$$\int_{-\infty}^{+\infty} h(z) \exp\left(-z^2\right) dz, \tag{31}$$

with a weighted sum, namely

$$\int_{-\infty}^{+\infty} h(z) \exp\left(-z^2\right) dz \approx \sum_{l=1}^{Q} w_l h(z_l). \tag{32}$$

Q is the order of the approximation, the z_l are the zeros of the Qth order Hermite polynomial and the w_l are corresponding weights. The nodes (or quadrature points) z_l and the weights w_l are tabulated in Abramowitz and Stegun (1972) (page 924). The quadrature points used in (32) do not depend on h. As such, it is possible that only very few nodes lie in the region where most of the mass of h is, which would lead to poor approximations.

Adaptive Gauss–Hermite quadrature. With an *adaptive* Gauss-Hermite quadrature rule the nodes are rescaled and shifted such that the integrand is sampled in a suitable range. Assume $h(z)\phi(z;0,1)$ is unimodal and consider the numerical integration of $\int_{-\infty}^{+\infty} h(z)\phi(z;0,1)dz$. Let $\hat{\mu}$ and $\hat{\nu}$ be

$$\hat{\mu} = \text{mode}\left[h(z)\phi(z;0,1)\right] \quad \text{and} \quad \hat{\nu}^2 = \left[-\frac{\partial^2}{\partial z^2} \ln\left(h(z)\phi(z;0,1)\right)\Big|_{z=\hat{\mu}}\right]^{-1}. \tag{33}$$

Acting as if $h(z)\phi(z;0,1)$ were a Gaussian density, $\hat{\mu}$ and $\hat{\nu}$ would be the mean and variance of this density. The quadrature points in the adaptive procedure, $z_l^\star$, are centered at $\hat{\mu}$ with spread determined by $\hat{\nu}$, namely

$$z_l^\star = \hat{\mu} + \sqrt{2}\hat{\nu} z_l \tag{34}$$

with $(l = 1, \ldots, Q)$. Now rewrite $\int_{-\infty}^{+\infty} h(z)\phi(z;0,1)dz$ as

$$\int_{-\infty}^{+\infty} \frac{h(z)\phi(z;0,1)}{\phi(z;\mu,\nu)} \phi(z;\mu,\nu) dz, \tag{35}$$

where $\phi(z;\mu,\nu)$ is the Gaussian density function with mean μ and variance ν^2. Using simple manipulations it is easy to see that for a suitably regular function v

$$\begin{aligned}\int_{-\infty}^{+\infty} v(z)\phi(z;\mu,\nu)dz &= \int_{-\infty}^{+\infty} v(z)(2\pi\nu^2)^{-1/2}\exp\left(-\frac{1}{2}\left(\frac{z-\mu}{\nu}\right)^2\right)dz \\ &= \int_{-\infty}^{+\infty}\frac{v(\mu+\sqrt{2}\nu z)}{\sqrt{\pi}}\exp\left(-z^2\right)dz \\ &\approx \sum_{l=1}^{Q}\frac{v(\mu+\sqrt{2}\nu z_l)}{\sqrt{\pi}}w_l. \end{aligned} \tag{36}$$

Using $\frac{h(z)\phi(z;0,1)}{\phi(z;\mu,\nu)}$ instead of $v(z)$ and replacing μ and ν with their estimates from (33), results in the following quadrature formula

$$\begin{aligned}\int_{-\infty}^{+\infty} h(z)\phi(z;0,1)dz &\approx \sqrt{2}\hat{\nu}\sum_{l=1}^{Q} w_l\exp\left(z_l^2\right)\phi(z_l^\star;0,1)h(z_l^\star) \\ &= \sum_{l=1}^{Q} w_l^\star h(z_l^\star), \end{aligned} \tag{37}$$

with adaptive weights $w_l^\star := \sqrt{2}\hat{\nu}w_l\exp\left(z_l^2\right)\phi(z_l^\star;0,1)$. (37) is an *adaptive* Gauss-Hermite quadrature formula.

Link with Laplace approximation. We illustrate the connection between the Laplace approximation (from Section 3.3.1) and adaptive Gauss–Hermite quadrature with a single node. Indeed, when $Q = 1$ (i.e. the case of a single node), $z_1 = 0$ (from the Hermite polynomial) and $w_1 = 1$. The corresponding adaptive node and weight are $z_1^\star = \hat{\mu}$ and $w_1^\star = \sqrt{2}\hat{\nu}\phi(z_1^\star;0,1)$. The adaptive GH quadrature formula then becomes

$$\begin{aligned}&\int h(z)\phi(z;0,1)dz \approx \sqrt{2}\hat{\nu}\exp\left(\log\left(\phi(z_1^\star;0,1)h(z_1^\star)\right)\right) \\ \propto\ & (2\pi)^{1/2}\underbrace{\left|-\frac{\partial^2}{\partial z^2}\log\left(h(z)\phi(z;0,1)\right)\Big|_{z=\hat{\mu}}\right|^{-1/2}}_{\hat{\nu}}\exp\left\{\log\left(\phi(\hat{\mu};0,1)h(\hat{\mu})\right)\right\},\end{aligned} \tag{38}$$

where $\hat{\mu} = z_1^\star$ maximizes $h(z)\phi(z;0,1)$. This corresponds with the Laplace formula from (18).

Adaptive Gauss–Hermite quadrature for GLMMs. We describe the case of a GLMM with a single, normally distributed random effect $u_i \sim N(0,\sigma^2)$ for each cluster i. The use of adaptive Gauss–Hermite quadrature with GLMMs starts from determining the posterior mode of u_i. Since this posterior distribution depends on unknown fixed effects and variance parameters, we replace the unknown $\boldsymbol{\beta}$, ϕ and σ with their current estimates: $\hat{\boldsymbol{\beta}}^{(c)}$, $\hat{\phi}^{(c)}$ and $\hat{\sigma}^{(c)}$. Using these current estimates $\hat{u}_i$ maximizes

$$f(\boldsymbol{y}_i|u_i)f(u_i|\hat{\sigma}^{(c)}), \tag{39}$$

which is proportional to the posterior density of u_i, given $\boldsymbol{y}_i$

$$\begin{aligned} f(u_i|\boldsymbol{y}_i) &= \frac{f(\boldsymbol{y}_i|u_i)f(u_i|\hat{\sigma}^{(c)})}{\int f(\boldsymbol{y}_i|u_i)f(u_i|\hat{\sigma}^{(c)})du_i} \\ &\propto f(\boldsymbol{y}_i|u_i)f(u_i|\hat{\sigma}^{(c)}). \end{aligned} \tag{40}$$

Therefore $\hat{u}_i$ is the posterior mode of u_i. We also determine (numerically) ${\hat{\nu}_i}^2$ as

$${\hat{\nu}_i}^2 = \left[-\frac{\partial^2}{\partial u_i^2} \ln \left(f(\boldsymbol{y}_i|u_i)f(u_i|\hat{\sigma}^{(c)}) \right) \Big|_{u_i=\hat{u}_i} \right]^{-1}. \tag{41}$$

Using an adaptive Gauss–Hermite quadrature rule we approximate the likelihood contribution of cluster i as follows (with $\delta_i := \sigma^{-1}u_i \sim N(0,1)$)

$$\begin{aligned} \int f_{\boldsymbol{Y}|U}(\boldsymbol{y}_i|u_i)f_U(u_i)du_i &= \int f_{\boldsymbol{Y}|U}(\boldsymbol{y}_i|\delta_i)\phi(\delta_i|0,1)d\delta_i \\ &= \int \left(\prod_{j=1}^{n_i} f_{Y|U}(y_{ij}|\delta_i) \right) \phi(\delta_i|0,1)d\delta_i \\ &\approx \sum_{l=1}^{Q} w_l^\star \left(\prod_{j=1}^{n_i} f_{Y|U}(y_{ij}|z_l^\star) \right), \end{aligned} \tag{42}$$

with adaptive weights $w_l^\star = \sqrt{2}\hat{\nu}_i w_l \exp(z_l^2)\phi(z_l^\star;0,1)$ and $z_l^\star = \hat{\delta}_i + \sqrt{2}\hat{\nu}_i z_l$. In this expression the linear predictor corresponding with $f_{Y|U}(y_{ij}|\delta_i)$ and $f_{Y|U}(y_{ij}|z_l^\star)$, respectively, is $\boldsymbol{x}_{ij}'\boldsymbol{\beta} + \sigma\delta_i$ and $\boldsymbol{x}_{ij}'\boldsymbol{\beta} + \sigma z_l^\star$. Multiplying (42) over all clusters i leads to the total likelihood. Maximizing the latter over the fixed effects regression parameters, the dispersion parameter and the variance components leads to updated parameter estimates $\hat{\beta}^{(c+1)}$, $\phi^{(c+1)}$ and $\hat{\sigma}^{(c+1)}$. We predict the cluster–specific random effects with the posterior modes from (39).

3.4 Pros and cons of various estimation methods for GLMMs

Laplace and PQL methods (as described in Section 3.3.1 and 3.3.2) for estimation within GLMMs rely on quite a few approximations. Breslow and Lin (1995) and Lin and Breslow (1996) investigate settings in which PQL (which results in the iterative approach from Section 3.3.2) performs poorly, and discuss the limits of this approach. Based on this McCulloch and Searle (2001) decide *"We thus cannot recommend the use of simple PQL methods in practice."* (see McCulloch and Searle (2001), Chapter 10, page 283). Gauss–Hermite quadrature is more accurate than PQL but limited to GLMMs with a small number of nested random effects. It is not possible to handle a large number of random effects, crossed random effects or high levels of nesting with this approach. Moreover, Gauss–Hermite quadrature is explicitly designed for normally distributed random effects, although other quadrature formulas exist (not discussed here).

The (Monte Carlo) EM algorithm and simulated maximum likelihood or Monte Carlo integration (see McCulloch and Searle (2001), Chapter 10, or Tuerlinckx et al. (2006)) are alternative methods for estimation with GLMMs.

We discuss a Bayesian implementation of (G)LMMs in Section 5. This is a way to circumvent the estimation problems discussed above.

3.5 Statistical inference with GLMMs

The general ideas on statistical inference with LMMs carry over to GLMMs where fitting is based on maximum likelihood principles. Wald, score and likelihood ratio tests ([LRT]) are available for hypothesis testing with fixed effects parameters, as well as variance components. However, closed–form expressions, for example for the covariance matrix of $\hat{\boldsymbol{\beta}}$, are no longer available. Numerical evaluation of the inverse Fisher information matrix is required for precision estimates. When using the PL method as described in Section 3.3.2, the original likelihood expression should be used in a LRT, and not the likelihood of the LMM that is specified for the pseudo–data. As with LMMs, testing the necessity of a random effect is problematic, since the corresponding null hypothesis constrains the variance of the random effect to the boundary of its parameter space. With respect to inference with (G)LMMs a Bayesian analysis has some additional features, see Section 5 for discussion.

4 Non–linear mixed models

LMMs and GLMMs model the mean (in LMMs) or a transformation of the (conditional) mean (in GLMMs) as *linear* in the fixed effects parameters $\boldsymbol{\beta}$ and the random effects $\boldsymbol{u}$. Non–linear mixed models ([NLMM]) release the concept of linear predictors. In a NLMM the conditional distribution of Y_{ij} (being the jth response on cluster i), given $\boldsymbol{u}_i$, belongs to the exponential family with mean structure

$$E[Y_{ij}|\boldsymbol{u}_i] \;=\; h(\boldsymbol{x}_{ij}, \boldsymbol{\beta}, \boldsymbol{z}_{ij}, \boldsymbol{u}_i), \tag{43}$$

where $h(.)$ is an arbitrary function of covariates, parameters and random effects. A distributional assumption for the random effects completes the model assumptions; typically $\boldsymbol{u}_i \sim N(\boldsymbol{0}, \boldsymbol{D})$. GLMMs are therefore a subclass of the general class of NLMMs. (Adaptive) Gauss–Hermite quadrature is available for ML estimation within NLMMs. A fully Bayesian analysis is an alternative approach.

5 Bayesian approach to (L,GL,NL)MMs

The presence of random effects is an essential feature in the hierarchical model formulation of a mixed model. A link with Bayesian statistics is then straightforward, since the random effects have explicit distributional assumptions. In addition to the distribution of the random effects $\boldsymbol{u}$ and the distributional framework for the response $\boldsymbol{y}$, a Bayesian analysis requires prior distributions for $\boldsymbol{\beta}$, (ϕ in GLMMs) and $\boldsymbol{D}$. Inference is based on simulated samples from the posterior distribution of the parameters, which is (with m clusters)

$$\begin{aligned} & f(\boldsymbol{\beta}, \boldsymbol{D}, \phi, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_m | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_m) \\ \propto \; & \prod_{i=1}^{m} f_i(\boldsymbol{y}_i | \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_m) \cdot \prod_{i=1}^{m} f(\boldsymbol{u}_i | \boldsymbol{D}) \cdot f(\boldsymbol{D}) \cdot f(\boldsymbol{\beta}) \cdot f(\phi). \end{aligned} \tag{44}$$

We refer to Chapters XXX and XXX on Bayesian concepts and regression models for an overview of useful concepts and simulation methods. For GLMMs in particular Zhao et al. (2006) and the references herein are a nice starting point for Bayesian (L,G,NL)MMs.

Bayesian multilevel models have some very nice features. As discussed in Chapter XXX on LMMs (see Section XXX), precision estimates based on MLE require variance components estimates to be plugged in, and are therefore not able to account for all sources of randomness. A fully Bayesian approach, with a prior specified for each parameter (vector), solves this issue and provides a way to circumvent otherwise intractable calculations. The likelihood approximations discussed in Section 3.3 are replaced in a Bayesian analysis with general MCMC methodology for sampling from posterior distributions. This allows specification of more complex hierarchical model structures, such as the spatial structures in Chapter XXX or the 3–level count data models in Antonio et al. (2010). Moreover, the Bayesian methodology is not limited to Gaussian random effects. For predictive modeling in actuarial science Bayesian statistics is particularly useful for simulation from the posterior (predictive) distribution of quantities of interest, such as a policy's random effect or the number of claims in a future time period.

[Reference to Chapters on Bayesian statistics.]

6 Examples

6.1 Poisson regression for workers' compensation insurance frequencies

We analyze the data from Illustration XXX (see Chapter XXX) on claim counts reported by 133 occupation classes with respect to their workers' compensation insurance policy. Each occupation class is followed over 7 years. The response variable of interest is Count_{ij}, the number of claims registered per occupation class i during year j. To enable out–of–sample predictions, we split the data in a training (without Count_{i7}) versus validation set (the Count_{i7} observations). We remove observations with zero payroll. Models are estimated on the training set, and centering of covariate Year is applied. Since the data are claim counts, we investigate the use of Poisson regression models. Throughout our analysis we include $\log(\mathsf{Payroll}_{ij})$ as an offset in the regression models, since the number of accidents should be interpreted relative to the size of the risk class.

From the discussion in Section 3.3 we are aware of (at least) three ways to tackle the problem of likelihood optimization with GLMMs. Correspondingly, multiple `R` packages are available for calibrating GLMMs to data. We illustrate hereafter the packages `lme4`, `glmmML` and the function `glmmPQL` from library `MASS` for likelihood–based estimation of a Poisson model with random effects. The illustration ends with a demonstration of a Bayesian analysis of this Poisson regression model.

Complete pooling. Similar to our approach in Chapter XXX, we start with a *'complete pooling model'*, ignoring the clustering of data in occupation classes. This is a simple Poisson regression model, with an overall intercept β_0 and an overall slope β_1 for the effect of Year.

$$\begin{aligned} \mathsf{Count}_{ij} &\sim \text{POI}(\mathsf{Payroll}_{ij} \cdot \lambda_{ij}) \\ \lambda_{ij} &= \exp(\beta_0 + \beta_1 \cdot \mathsf{Year}_{ij}). \end{aligned} \tag{45}$$

A Poisson regression is an example of a Generalized Linear Model (see Chapter XXX) for which the `glm` function in `R` is available.

```
> fitglm.CP <- glm(count~yearcentr, offset=log(payroll),family=poisson,
                                  data=wcFit)
> summary(fitglm.CP)

Call:
glm(formula = count ~ yearcentr, family = poisson, data = wcFit,
    offset = log(payroll))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-26.8194   -1.0449    0.2456    2.3197   18.1740

Coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept) -3.702172   0.008648 -428.105   <2e-16 ***
yearcentr   -0.010155   0.005098   -1.992   0.0464 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 12274  on 766  degrees of freedom
Residual deviance: 12270  on 765  degrees of freedom
AIC: 14904

Number of Fisher Scoring iterations: 5
```

The estimate $\hat{\beta}_0$ for the intercept is -3.702 (with s.e. 0.00865) and $\hat{\beta}_1 = -0.0102$ (with s.e. 0.00510).

No pooling. We continue the analysis with a fixed effects Poisson model that specifies an occupation class specific intercept, say $\beta_{0,i}$, for each of the 113 occupation classes in the data set, as well as a global, fixed Year effect. The intercepts $\beta_{0,i}$ are unknown, but fixed. We fit the model in R with the `glm` function, identifying the occupation class as a factor variable.

$$\begin{aligned} \mathsf{Count}_{ij} &\sim \text{POI}(\mathsf{Payroll}_{ij} \cdot \lambda_{ij}) \\ \lambda_{ij} &= \exp\left(\beta_{0,i} + \beta_1 \cdot \mathsf{Year}_{ij}\right). \end{aligned} \tag{46}$$

```
> fitglm.NP <- glm(count~0+yearcentr+factor(riskclass), offset=log(payroll),
                             family=poisson(),data=wcFit)
> summary(fitglm.NP)
Call:
glm(formula = count ~ 0 + yearcentr + factor(riskclass), family = poisson(),
```

```
    data = wcFit, offset = log(payroll))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.2403  -0.8507  -0.1629   0.7186   7.1909

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
yearcentr             9.918e-03  5.157e-03   1.923 0.054448 .
factor(riskclass)1   -2.578e+00  2.425e-01 -10.630  < 2e-16 ***
factor(riskclass)2   -3.655e+00  4.082e-01  -8.952  < 2e-16 ***
factor(riskclass)3   -3.683e+00  1.374e-01 -26.810  < 2e-16 ***
factor(riskclass)4   -1.309e+01  2.103e+03  -0.006 0.995035
factor(riskclass)5   -2.737e+00  9.325e-02 -29.347  < 2e-16 ***
...
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 980297.4  on 767  degrees of freedom
Residual deviance:   1192.9  on 636  degrees of freedom
AIC: 4084.4

Number of Fisher Scoring iterations: 14
```

Comparing the deviance of the *'complete'* and the *'no pooling'* model results in a drop–in–deviance of 12,270-1192=11,078, with a difference in degrees of freedom of 129. In `R` the `anova` function is available for this comparison. With a p–value of $< 2.2 \cdot 10^{-16}$ the *'no pooling'* model outperforms the *'complete pooling'* model.

```
> anova(fitglm.CP,fitglm.NP,test="Chisq")
Analysis of Deviance Table

Model 1: count ~ yearcentr
Model 2: count ~ 0 + yearcentr + factor(riskclass)
  Resid. Df Resid. Dev  Df Deviance  Pr(>Chi)
1       765    12270.4
2       636     1192.9 129    11078 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

GLMMs: random intercepts, Laplace approximation with `lme4`. We investigate a Poisson regression model with random occupation class specific intercepts as a meaningful alternative for the *'no pooling'* model. The model formulation is

$$\begin{aligned} \mathsf{Count}_{ij}|u_{i,0} &\sim \text{POI}(\mathsf{Payroll}_{ij} \cdot (\lambda_{ij}|u_{i,0})) \\ \lambda_{ij}|u_{i,0} &= \exp\left(\beta_0 + u_{i,0} + \beta_1 \cdot \mathsf{Year}_{ij}\right) \\ u_{i,0} &\sim N(0, \sigma_u^2). \end{aligned} \tag{47}$$

We first fit this random intercepts model with the `lmer` (or: `glmer`) function from the `R` library `lme4`. By default `lmer` is based on Laplace approximation (see Section 3.3.1) to optimize the Poisson likelihood (though it does not ignore the last term in 23, see `http://lme4.r-forge.r-project.org/book/`). Adaptive Gauss–Hermite quadrature is also available within the `lme4` package (see infra) [3].

```
> hlm1 <- glmer(count ~ (1|riskclass)+yearcentr+offset(log(payroll)),
+           family=poisson(link="log"), data=wcFit)
> print(hlm1)
Generalized linear mixed model fit by the Laplace approximation
Formula: count ~ (1 | riskclass) + yearcentr + offset(log(payroll))
   Data: wcFit
  AIC  BIC logLik deviance
 1771 1785 -882.6     1765
Random effects:
 Groups    Name        Variance Std.Dev.
 riskclass (Intercept) 0.80475  0.89708
Number of obs: 767, groups: riskclass, 130

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.562125   0.083489  -42.67   <2e-16 ***
yearcentr    0.009730   0.005156    1.89   0.0592 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Correlation of Fixed Effects:
          (Intr)
yearcentr -0.010
```

From the `R` output above we conclude $\hat{\beta}_0 = -3.5621$ (with s.e. 0.0835), $\hat{\beta}_1 = 0.00973$ (with s.e. 0.00516), and $\hat{\sigma}_u^2 = 0.805$ the estimate of the variance of random intercepts. Extracting the estimates of fixed and random effects, as well as prediction intervals for the random intercepts, goes as follows

[3] `glmmmML` package is another `R` package for Laplace approximation and Gauss–Hermite quadrature for Binomial and Poisson random effects models, see the book's supporting web page for sample code.

```
> ## get fixed effects
> fixef(hlm1)
 (Intercept)    yearcentr
-3.562124594  0.009730364

> ## get random intercepts
> int <- ranef(hlm1)$riskclass

> ## get prediction intervals for r.e.'s
> str(rr1 <- ranef(hlm1, condVar = TRUE))
> # s.e. for 'riskclass' r.e.
> my.se.risk = sqrt(as.numeric(attributes(rr1$riskclass)$postVar))
> # get prediction intervals for random intercepts (per riskclass)
> lower.risk <- rr1$riskclass[[1]]-1.96*my.se.risk
> upper.risk <- rr1$riskclass[[1]]+1.96*my.se.risk
> int.risk <- cbind((lower.risk),(rr1$riskclass[[1]]),(upper.risk))
> colnames(int.risk)  <- c("Lower","Estimate R.E.","Upper")
# you can use these to create error bar plots
> int.risk[1:5,]
          Lower Estimate R.E.     Upper
[1,]  0.4407844  0.9146787935 1.3885732
[2,] -0.8002651 -0.0767152172 0.6468347
[3,] -0.3834920 -0.1177250934 0.1480418
[4,] -1.7583808 -0.0009170246 1.7565468
[5,]  0.6340478  0.8166379517 0.9992281
>
> ## variance components
> VarCorr(hlm1)
$riskclass
            (Intercept)
(Intercept)   0.8047458
attr(,"stddev")
(Intercept)
  0.8970762
attr(,"correlation")
            (Intercept)
(Intercept)           1

attr(,"sc")
[1] NA
```

We are now ready to compare model (46) (*'no pooling'*) to (47) (random intercepts). The left panel in Figure 2 shows the intercepts and corresponding error bars from the *'no pooling'* model, together with one standard error, against the total size of the occupation class (i.e. $\sum_j \mathsf{Payroll}_{ij}$). Figure 2 (right) shows the point predictions of the random intercepts. To create this plot we refit the random effects model and do not include an intercept [4]. The blue dashed line in the Figure is $y = -3.702$, the overall intercept from the *complete pooling* model.

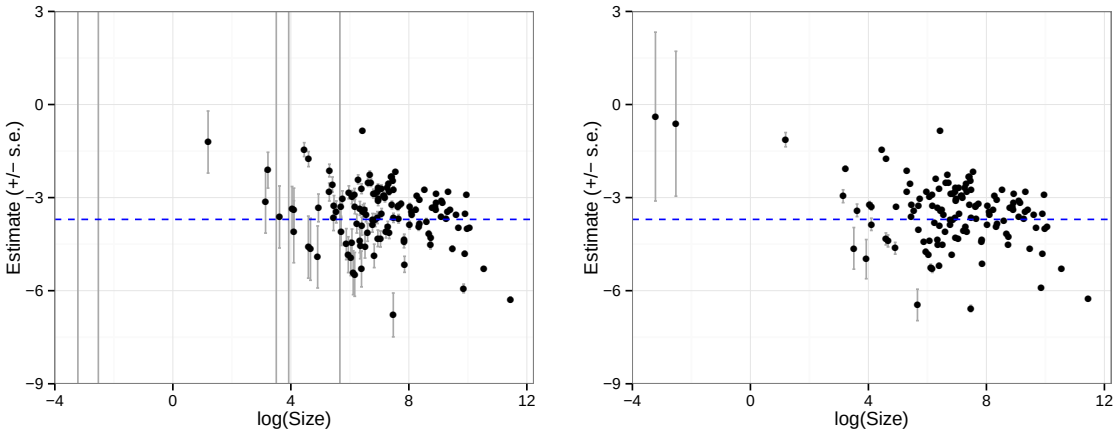

Figure 2: *Point estimates for occupation class specific intercepts, plus/minus one standard error. Results from no pooling approach (left) and Poisson random intercept model (right). The dashed line is $y = -3.702$, i.e. the overall intercept from the complete pooling model.*

As discussed in the Introduction of Chapter XXX the *'no pooling'* model results in unreasonable estimates for certain occupation classes. The output printed below compares the size, random intercept estimate, fixed intercept estimate and corresponding standard errors for a selection of occupation classes.

```
## occupation class 122 (our numbering)
# random intercept model
num  size     lower  estimate    upper  logsize
122 33.32 -6.228015 -4.635567 -3.04312  3.506158
# no pooling model
num  size     lower  estimate    upper  logsize
122 33.32 -854.0756 -17.96726 818.1411  3.506158
# data for this class (i.e. zero claims on 33.32 payroll total)
riskclass year count    payroll
      122    1     0    3.28
      122    2     0    5.69
```

[4] As explained in Chapter XXX on LMMs, `lme4` evaluates the variance of the distribution of $\boldsymbol{u}|\boldsymbol{y}$, conditional on the maximum likelihood estimates for unknown parameters

```
      122    3     0    4.51
      122    4     0    4.80
      122    5     0    9.07
      122    6     0    5.97
## occupation class (our numbering)
# random intercepts model
num  size     lower  estimate     upper logsize
61 23.16 -3.840979 -2.955994 -2.071008  3.142427
# no pooling model
num  size     lower  estimate     upper logsize
61 23.16 -4.144029 -3.144028 -2.144028  3.142427
# data for this class (i.e. 1 claim on 23.26 payroll total)
riskclass year count payroll
 61    1     0    3.12
 61    2     0    3.68
 61    3     0    3.76
 61    4     0    3.83
 61    5     1    4.99
 61    6     0    3.78
## occupation class (our numbering)
# random intercepts model
num  size     lower       estimate          upper     logsize
52   0.08    -3.5900335  -0.6187492     2.3525350    -2.525729
# no pooling model
num  size     lower       estimate          upper      logsize
52   0.08 -2117.1442456 -13.7817186     2089.5808085   -2.525729
# data for this class (i.e. 0 claims on 0.08 payroll total)
riskclass year count payroll
52          4     0     0.08
```

GLMMs: random intercepts, adaptive Gauss–Hermite quadrature with `lme4`. Adaptive Gauss–Hermite quadrature (see Section 3.3.3) is available within the `lme4` package. We estimate the random intercepts model from (47) again this technique, and opt for 15 quadrature points (see `nAGQ=15`).

```
> hlm2 <- lmer(count ~ (1|riskclass)+yearcentr+offset(log(payroll)),
+           family=poisson(), data=wcFit,nAGQ=15)
> print(hlm2)
Generalized linear mixed model fit by the adaptive Gaussian Hermite approximation
Formula: count ~ (1 | riskclass) + yearcentr + offset(log(payroll))
   Data: wcFit
  AIC  BIC logLik deviance
 1771 1785 -882.4     1765
Random effects:
```

```
 Groups    Name        Variance Std.Dev.
 riskclass (Intercept) 0.80733  0.89851
Number of obs: 767, groups: riskclass, 130

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.561974   0.083614  -42.60   <2e-16 ***
yearcentr    0.009731   0.005156    1.89   0.0591 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Correlation of Fixed Effects:
          (Intr)
yearcentr -0.010
```

The parameter estimates and standard errors as obtained with GH quadrature are very close to those obtained with Laplace approximation (see the results of `hlm1`). Changing the number of quadrature points has very minor impact on the results.

GLMMs: random intercepts, approximating the data with `glmmmPQL`. To illustrate the approach from Section 3.3.2, i.e. repetitive fits of a linear mixed model to pseudo–data, the function `glmmmPQL` from `library(MASS)` is available in `R`. For the Poisson random intercepts model from (47) convergence is reached in 10 iterations. Parameter estimates and corresponding standard errors are printed below. They are different from the results obtained with `lme4`, though close. Based on the discussion in Section 3.4 however, we prefer the use of adaptive Gauss–Hermite quadrature, whenever possible for the model under consideration.

```
> library(MASS)
> PQL1 <- glmmPQL(count ~ yearcentr + offset(log(payroll)),
                random = ~ 1 | riskclass,family = poisson, data = wcFit)
iteration 1
iteration 2
iteration 3
iteration 4
iteration 5
iteration 6
iteration 7
iteration 8
iteration 9
iteration 10
> summary(PQL1)
Linear mixed-effects model fit by maximum likelihood
 Data: wcFit
```

```
  AIC BIC logLik
   NA  NA     NA

Random effects:
 Formula: ~1 | riskclass
        (Intercept) Residual
StdDev:   0.9290198  1.50974

Variance function:
 Structure: fixed weights
 Formula: ~invwt
Fixed effects: count ~ yearcentr + offset(log(payroll))
                Value  Std.Error  DF   t-value p-value
(Intercept) -3.489754 0.08911097 636 -39.16189  0.0000
yearcentr    0.009496 0.00656277 636   1.44688  0.1484
 Correlation:
          (Intr)
yearcentr -0.012

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-2.5749914 -0.5294022 -0.1518360  0.4497736 12.6121268

Number of Observations: 767
Number of Groups: 130
```

GLMMs: random intercepts, a Bayesian approach. Finally, we present a Bayesian analysis of the random intercepts model in (47). We analyze this example using `WinBUGS` and its interface with `R`, namely the function `bugs` from library `BRugs`. On the book's support page we also demonstrate the use of `R` package `glmmBUGS`. In `WinBUGS` the random intercepts model in (47) is coded as follows

```
model;
{
for(i in 1:895){
mu[i] <- payroll[i]*exp(beta1*yearcentred[i]+b[riskclass[i]])
count[i] ~ dpois(mu[i])
}
#specify distribution for fixed effects
beta0 ~ dnorm(0,0.0001)
beta1 ~ dnorm(0,0.0001)
#specify distribution for random effects
for(i in 1:133){
b[i] ~ dnorm(beta0,taub)
```

```
}
taub ~ dgamma(0.01,0.01)
sigma2b <- 1/taub
}
```

where we use normal $N(0, 10^{-4})$ priors for β_0 and β_1, and a $\Gamma(0.01, 0.01)$ prior for $(\sigma_u^2)^{-1}$. The posterior densities of these parameters are illustrated in Figure 3.

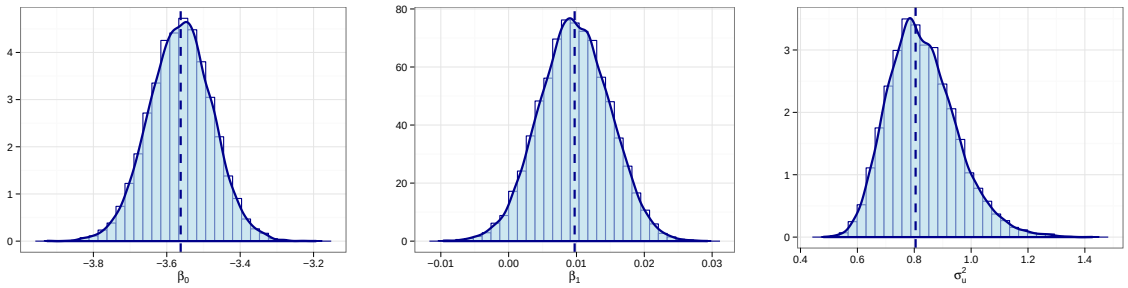

Figure 3: *Posterior simulations for parameters used in* (47) *(from left to right:* β_0*,* β_1 *and* σ_u^2*), workers' compensation insurance (frequencies). Results are based on 2 chains, 50,000 simulations each, thinning factor of 5 and burn–in of 2,000 simulations.*

With respect to predictive modeling, a Bayesian approach is most useful, since it provides the full predictive distribution of variables of interest (here: $\texttt{Count}_{i7}$). We illustrate this in Figure 4 for a selection of risk classes. Histograms are based on 50,000 simulations from the relevant predictive distribution (using model (47)). For each risk class the observed number of claims is indicated, as well as the point prediction obtained with a frequentist approach, using Laplace approximation from `(g)lmer`.

6.1.1 Tweedie compound Poisson mixed models

For the Tweedie compound Poisson mixed models with a variance function $V(\mu) = \mu^p$ for some $p \in (1, 2)$, we seek to estimate the unknown variance function, i.e., the index parameter p from the data along with the fixed effects and the variance component. This parameter has a significant impact on hypothesis tests and predictive uncertainty measures (Davidian and Carroll, 1987; Peters et al., 2009; Zhang, 2012), which is of special interest to the insurance industry. For example, if a Tweedie compound Poisson GLM is exploited in loss reserving modeling, the uncertainty measures of the predicted outstanding liability will be substantially influenced by the choice/estimation of the index parameter.

One approach in estimating the variance function is using the profile likelihood (Cox and Reid, 1987). For the compound Poisson distribution, such an approach must be implemented based on the true likelihood rather than the quasi-likelihood. It is well known that the basic quasi-likelihood method, and hence the PQL method introduced above is not equipped to estimate the unknown variance function. Its natural extension, the extended quasi-likelihood (Nelder and Pregibon, 1987), however, is also not well suited

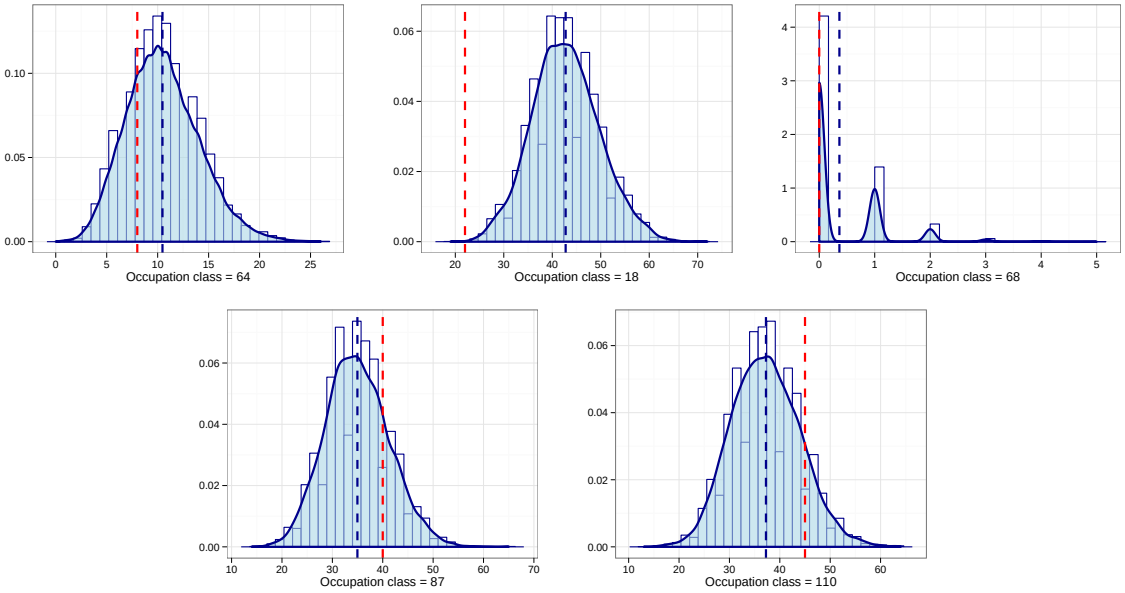

Figure 4: *Posterior predictive simulations for the number of claims in year 7 for a selection of risk classes. Simulations are based on Bayesian analysis of* (47)*, using 2 chains, 50,000 simulations each, thinning factor of 5 and burn–in of 2,000 simulations. The dashed blue line is the point prediction as obtained with* `(g)lmer`*, the dashed red line is the observed number of claims.*

to this task in that it involves a term $\log(V(y))$ which becomes infinite for $y = 0$. Its implementation therefore requires adding a small positive constant to the observed zeros which, unfortunately, is highly influential on parameter estimation (see Zhang (2012)).

Likelihood-based methods, namely, the adaptive Gauss-Hermite quadrature method and the Laplace approximation (with one quadrature point), must be used to enable data-driven estimation of the index parameter. Yet, a complicating factor is that the compound Poisson distribution has an intractable density function. When performing maximum likelihood estimation, we must rely on numerical methods to approximate the density function, that is, the conditional distribution of the data given the random effects. Such numerical methods that allow fast and accurate evaluation of the compound Poisson density function are provided in Dunn and Smyth (2005, 2008). Similarly to the above, the approximated likelihood is then optimized numerically to produce parameter estimates, including the maximum likelihood estimate for p.

Both likelihood-based and Bayesian methods for the Tweedie compound Poisson mixed model have been implemented in the `R` package `cplm`.

References

Abramowitz, M. and Stegun, I. (1972). *Handbook of mathematical functions: with formulas, graphs and mathematical tables*. Dover, New York.

Antonio, K., Frees, E., and Valdez, E. (2010). A multilevel analysis of intercompany claim counts. *ASTIN Bulletin: The Journal of the International Actuarial Association*, 40(1):151–177.

Antonio, K. and Valdez, E. (2012). Statistical aspects of *a priori* and *a posteriori* risk classification in insurance. *Advances in Statistical Analysis*, 96(2):187–224.

Breslow, N. and Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

Breslow, N. and Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika*, 82:81–91.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society, Series B*, 49:1–39.

Davidian, M. and Carroll, R. J. (1987). Variance function estimation. *Journal of the American Statistical Association*, 82:1079–1091.

Denuit, M., Maréchal, X., Pitrebois, S., and Walhin, J.-F. (2007). *Actuarial Modelling Of Claim Counts: Risk Classification, Credibility and Bonus-Malus Scales*. Wiley.

Dunn, P. K. and Smyth, G. K. (2005). Series evaluation of tweedie exponential dispersion models densities. *Statistics and Computing*, 15:267–280.

Dunn, P. K. and Smyth, G. K. (2008). Evaluation of tweedie exponential dispersion model densities by fourier inversion. *Statistics and Computing*, 18:73–86.

Haberman, S. and Renshaw, A. (1996). Generalized linear models and actuarial science. *The Statistician*, 45(4):407–436.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Lin, X. and Breslow, N. (1996). Analysis of correlated binomial data in logistic-normal models. *Journal of Statistical Computation and Simulation*, 55:133–146.

Liu, Q. and Pierce, D. (1994). A note on Gauss-Hermite quadrature. *Biometrika*, 81(3):624–629.

McCulloch, C. and Searle, S. (2001). *Generalized, Linear and Mixed Models*. Wiley Series in Probability and Statistics, Wiley, New York.

Molenberghs, G. and Verbeke, G. (2005). *Models for discrete longitudinal data*. Springer Series In Statistics, Springer, New York.

Nelder, J. and Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika*, 74:221–232.

Nelder, J. and Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135:370–384.

Peters, G. W., Shevchenko, P. V., and Wüthrich, M. V. (2009). Model uncertainty in claims reserving within tweedie's compound poisson models. *ASTIN Bulletin*, 39:1–33.

Pinheiro, J. and Bates, D. (2000). *Mixed Effects Models in S and S-Plus*. Springer.

Purcaru, O., Guillén, M., and Denuit, M. (2004). Linear credibility models based on time series for claim counts. *Belgian Actuarial Bulletin*, 4(1):62–74.

Tierny, L. and Kadane, J. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81:82–86.

Tuerlinckx, F., Rijmen, F., Verbeke, G., and Boeck, P. D. (2006). Statistical inference in generalized linear mixed models: a review. *British Journal of Mathematical and Statistical Psychology*, 59:225–255.

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 48:233–243.

Yip, K. C. H. and Yau, K. K. W. (2005). On modeling claim frequency data in general insurance with extra zeros. *Insurance: Mathematics and Economics*, 36:153–163.

Zhang, Y. (2012). Likelihood-based and bayesian methods for tweedie compound poisson linear mixed models. *Statistics and Computing*. forthcoming.

Zhao, Y., Staudenmayer, J., Coull, B., and Wand, M. (2006). General design bayesian generalized linear mixed models. *Statistical Science*, 21:35–51.

FACULTY OF ECONOMICS AND BUSINESS
Naamsestraat 69 bus 3500
3000 LEUVEN, BELGIË
tel. + 32 16 32 66 12
fax + 32 16 32 67 91
info@econ.kuleuven.be
www.econ.kuleuven.be